sequences. The second is the periodic property of the DNA sequences. The third is that amount of information of the sequences. By using this method, we classify the nature sequences and artifical sequences. At last, we analyze the characteristic in this model and consider the generalization of this model.

关于DNA 序列分类问题的模型

冯 涛, 康喆雯, 韩小军
指导老师: 贺明峰

(大连理工大学, 大连 116024)

编者按: 本文以统计方法提取样本特征, 以之作为BP 神经网络的输入, 用MATLAB 中相应算法进行训练. 然后用于解决本分类问题, 得到了较准确的结果. 本文提取特征时考虑较为全面, 在此基础上正确地运用了神经网络方法, 发挥了神经网络适用于非线性问题、具有自适应能力的优点. 思路清楚, 文字简练.

摘要: 本文提出了一种将人工神经元网络用于DNA 分类的方法. 作者首先应用概率统计的方法对 20 个已知类别的人工DNA 序列进行特征提取, 形成DNA 序列的特征向量, 并将之作为样本输入BP 神经网络进行学习. 作者应用了MATLAB 软件包中的Neural Network Toolbox (神经网络工具箱) 中的反向传播(Back propagation BP) 算法来训练神经网络. 在本文中, 作者构造了两个三层BP 神经网络, 将提取的DNA 特征向量集作为样本分别输入这两个网络进行学习. 通过训练后, 将 20 个未分类的人工序列样本和 182 个自然序列样本提取特征形成特征向量并输入两个网络进行分类. 结果表明: 本文中提出的分类方法能够以很高的正确率和精度对DNA 序列进行分类, 将人工神经元网络用于DNA 序列分类是完全可行的.

1 问题重述(略)

DNA 序列由四个碱基A、T、C、G 按一定规律排列而成. 已知所给人工序列 1- 10 属于A 类, 11- 20 属于B 类. 本题中, 我们的主要工作有两个:

1) 提取A、B 两类特征;

2) 以所提取A、B 两类特征为依据, 把 20 个人工序列及 182 个自然序列分为A、B 两类(可能存在同时不具有A、B 两类特征, 不能归为A、B 中任一类的序列).

在本题中, 先以序列 1- 20 为依据, 提取出A、B 两类序列的统计特征, 然后运用神经网络中的BP 网络对未知序列进行了分类识别.

2 模型建立的理论依据

神经网络是近年来发展的一种大规模并行分布处理的非线性系统[1], 其主要特点有:

1) 能以任意精度逼近任意给定连续的非线性函数;

2) 对复杂不确定问题具有自适应和自学习能力;

3) 具有较强的容错能力和信息综合能力, 能同时处理定量和定性的信息, 能很好地协调多种输入信息的关系.

传统的分类识别方法, 对于一般非线性系统的识别很困难, 而神经网络却为此提供了一

个强有力的工具　它实质上是选择了一个适当的神经网络模型来逼近实际系统　目前, 在神经网络中应用最多的是BP 网络

对于具有 n 个输入节点,m 个输出节点的BP 网络, 输入到输出的关系可以看作是一个 n 维欧式空间到 m 维欧式空间的映射, $F: R^n \rightarrow R^m$, 这一映射是高度非线性映射　K. T. Funahashi 于 1989 年证明了这样的一个定理[2]: 如果BP 网络隐层节点可以根据问题的不同作相应的配置的话, 那么用三层的激励函数为双曲线正切型的BP 网络, 可以以任意精度逼近任意连续函数　这一定理保证了BP 网络在分类识别问题中的可用性

将复杂系统看作是一个黑箱, 以实测输入, 输出数据为学习样本, 送入BP 网络, 网络通过样本进行学习, 在学习过程中, 网络的权值不断地修改[3], 使输入到输出的映象逐渐与实际对象的特性相逼近, 但网络输出的整体误差 E 小于给定的标准时, 整个网络便模拟出实际系统的外部特性

实际分类识别问题中, 输入空间一般是多维欧式空间, 我们可以计算空间中点与点的欧式距离, 并根据这些距离知道哪些样本互相靠得近, 哪些样本相距甚远, 也就是说在输入空间中存在着一个距离度量, 只要输入模式接近于某个输出模式, 由于BP 网络所具有的联想记忆能力, 则网络的输出亦会接近学习样本的输出

3　模型的基本假设

1) 假设碱基序列的特征值包括以下两个内容: (1) 单个碱基在序列中的数量特征, 即 A, T, C, G 四种碱基在序列中的含量; (2) 特征碱基串在序列中的数量特征(包括双字符碱基串和三字符碱基串).

2) 由于给定的已知碱基序列是从DNA 全序列中随机截取出来的, 因此无法确定序列的起始位, 无法从序列中辨认出氨基酸　假设在对DNA 序列分类时, 是从碱基层次上进行分类, 而不是从氨基酸层次上分类

4　模型的建立与求解

4 1　提取A、B 两类的特征

经过计算, 我们提取出A、B 两类的统计特征(a)和(b), 具体方法如下:

特征(a): 单个字符出现的频率　特征(a)对应基本假设 1 中的第 1 条

对 1- 20 每个人工序列, 我们统计出单个字符A、T、C、G 出现的频率 Pi, $P_i = T_i/(S - M + 1)$, i= A, T, C, G

S 为序列长度,M 为字符长度(这里,M = 1), T_i 为每个序列中 i 出现的次数

序列 1- 20 特征(a)的数值如下: (略)

特征(b): 特征字符串出现的频率　特征(b)对应基本假设 1 中的第 2 条

通过对序列 1- 20 种A、T、C、G 四字母的不同组合(如两两组合, 三三组合, 四四组合)出现频率的分析, 可以知道: 对于双字符串和三字符串, 均出现了数种多次出现较有规律的组合形式, 而对于四四组合及更长的组合, 字符串重复出现的频率小, 分散度大, 未得出较有规律的组合方式　我们认为: 充分统计并分析序列 1- 20 种双字符串及三字符串出现的规律已能较为全面地认识序列中的局部相关性及A、B 两类的特征差异　因此, 只对序列 1- 20 种的双、三字符串进行统计分析, 找出特征双字符串, 特征三字符串

以下是以提取特征三字符串为例介绍统计算法:

第一步 确定各字符串的优先权重

三字符串共有 64 种可能排列方式, 对这些三字符串进行初次排列, 确定优先权重

以A 类序列 1 为例, aggcacggaa. . . . . gcttgg

1) 指针指向第一个字符 a, 向后数两个字符, 第一个出现的三字符串是 agg, 记录 agg

2) 指针向后移一个字符, 第二个出现的三字符串是 ggc

3) 以此类推, 记录到该序列中最后一个三字符串(tgg) (特别的, 如果相邻两个字符串完全相同, 只纪录一次).

同理可得序列 2- 10 种所有出现的三字符串, 最后把A 类中所有这些三字符串按其出现频率大小进行排序, 出现频率多的字符串优先权重就大

第二步 选出特征字符串, 对字符串进行二次排序, 找出特征字符串

仍以A 类序列 1 为例: aggcacggaa

1) 先考虑前 5 个字符, aggca, 其中包含了 3 个三字符串: agg, ggc, gca, 按第一步所得的三字符串优先权重的大小, 确定这 3 个字符串中有一个为特征字符串(如果 ggc 在前 10 个序列中出现的频率比 agg 和 gca 大, 那么在本例中就选 ggc, 而不考虑第一个字符 a).

2) 再把指针移至特征字符串后的第一个字符(本例中移向 a) 重复 (1) 操作. 以此类推, 直至找出A 类序列 1- 10 种所有特征字符串

我们采用分类统计的方法进行排序,B 类的操作方法同A 类

第三步 把A、B 两类的所有特征字符串进行排序, 计算出每个特征字符串在两类序列(1- 20) 中出现的总次数. 如果小于 5 次, 认为此字符串不能体现A、B 两类的特征差异, 不予考虑. 这样, 统计出 1- 20 中出现频率较大的特征三字符串(共 21 种), 他们在每个序列中出现的频率为: 3 * 该字符串在本序列中出现的次数/(S—M + 1), 这里,M = 3)

统计特征二字符串时, 采取类似的方法, 得出 15 个特征二字符串: 他们在每个序列中出现的频率为: 2 * 该字符串在本序列中出现的次数/(S—M + 1), 这里,M = 3).

4 2 网络输入与输出变量的选取及处理

选取网络的输入变量时, 如输入变量过少, 能引起建模不充分, 过多的输入变量会降低网络的学习速度, 延长收敛时间, 使模型的输入输出关系过于复杂. 结合本题的实际情况, 我们提出两套输入变量选取方案

方案 1 输入每个序列中单字符及特征三字符串出现的频率(共 25 个输入变量)

方案 2 输入每个序列中单字符及特征双字符串出现的频率(共 19 个输入变量)

如果要同时考虑单字符, 特征双、三字符串出现的频率共需 40 个输入变量, 模型过于复杂. 因此, 暂不考虑这种方案

规定: A 类序列的期望输出值为- 1,B 类为 1. 这样, 通过观察BP 网络的输出值, 可以直观地判断未知序列的类别

4 3 BP 网络的结构与参数

BP 网络的结构与参数决定着网络学习的效果和分类识别的精度. 其中, 输入、输出节点数由实际问题决定, 本题中输出节点为 1 个. 需要选择的是网络的激发函数, 隐层数及各层隐节点数

对方案 1、2, 各构造网络 1、2 与之相对应. 对于这两个网络, 均选用三层BP 网络, 各层

激发函数均为双曲线正切函数(函数值在- 1～ + 1 之间变化).

R. P L ippmann 研究中指出[4]: 对于任给K 个实数值样本, 有 $2K + 1$ 个隐节点的三层网络可以记忆它们, 这个隐单元的激发函数可以是任何渐近函数　基于这一结论, 我们根据样本集的规模, 选隐层节点数 $N = 5$, 这样可使网络有能力记忆全体样本, 不至于在学习过程中丢失前面的学习过的样本的信息

4 4　网络的训练及检验

在已知类别序列 1～ 20 中, 取A 类前 7 个序列(1～ 7)和B 类前 7 个序列(11～ 17)作为训练样本集 Strain, 序列 8～ 10、18～ 20 作为检验样本集 Stest 对网络 1: 25- 5- 1 及网络 19 - 5- 1 进行训练, 给定样本总体误差标准为 10- 5　当网络学习收敛于给定的标准后, 用检验样本集进行分类检验, 考察其分类识别的准确性　网络 1、2 的初始权值均为- 0. 2～ + 0. 2 之间的随机数　学习算法采用了两种改进措施相结合的BP 算法, 即变周期和变步长相结合的方法, 用以提高网络的收敛速度　在网络 1 开始训练时, 学习率 η 取 0. 9(网络 2 取 1. 0), 惯性系数 α 取 0. 6(网络 2 取 α 为 0. 7), 修正周期 T 取 10. 随着误差 E 的减少, 网络不断逼近对象的输出特性, 此时, 逐渐减少 η 及 α, 增大 T, 直至网络收敛于给定的标准　训练达到稳定时, 两个网络对训练样本集的学习速率曲线如图 1(a)和图 2(a) (略), 此时对检验样本的检验结果如图 1(b)和图 2(b) (略):

图 1(a)和图 2(a), 网络 1 进行了 303 步, 网络 2 进行了 241 步的学习后, 就达到了精度要求, 均学习速率较快, 效率较高

图 1(b)和图 2(b), 如果允许误差为 10%, 那么此时网络 1 对检验样本分类的准确性为 98. 3%, 网络 2 为 94. 7%, 命中率均为 100%, 我们将检验集加入到训练集中, 得组合集 Strain+ test　网络用此集进行学习　收敛后, 网络 1、2 可对未知序列进行分类识别了.

5　结果及分析

5. 1　对人工序列 21～ 40 的分类

我们应用MA TLAB 软件包中的神经网络工具箱(BP 网络)对未知序列进行分类　我们发现: 若以高于 0. 9 和低于- 0. 9 作为分类标准, 两个BP 网络的命中率相同, 但输出函数值不等, 网络 1 的输出值与期望值更接近　这种情况出现的原因是:

①网络 2 中输入变量较网络 1 少, 在样本集个数相同的情况下, 建模不够充分;

②双字符串的组合形式较三字符串少, 因此, 采用特征三字符串能能更好的体现序列中片段的相关性

经过反复训练、检验、分类, 我们发现: 网络 1 较网络 2 学习速度快, 对未知序列区分的精度更高, 因此, 认为网络 1 更优

在这里, 采用网络 1 的分类结果, 即: A 类: 22, 23, 25, 27, 29, 34, 35, 37, 39; B 类: 21, 24, 26, 28, 30, 31, 32, 33, 36, 38, 40.

5 2　对 182 个自然序列的分类

我们把 21～ 40 中已明确分类的序列加入到样本中, 重新对网络 1 进行训练, 直至达到误差 10^{-5}. 分别以高于 0, 0. 2, 0. 5 和低于 0, 0. 2, 0. 5 作为分类标准, 对 182 个自然序列的分类结果为: (略)

随着分类标准的变化, 分类率随之变化　采用 0 作为分类标准可把 182 个自然序列分开

6 模型的优缺点及改进方向

优点:

① 基因特征这种非线性系统很难用数学方程表达出来, 而且可利用的样本有限, 以至于传统的分类识别方法显得无效, 神经网络从其良好的学习功能和很强的非线性计算能力, 为分类提供了一种新方法;

② 传统的分类方法是一种模型驱动方法, 大部分统计模型基于线性回归, 而神经网络用数据驱动方式来解决分类问题, 它通过样本学习逼近实际系统模型的能力很强;

③ 由于BP 网络的信息分布性, 各输入变量对输出变量的影响在对样本学习时已自动记下, 并由整个网络的内部表达而表现出来, 从而省略了通常建模前所需的对各变量的相关分析;

④ BP 网络有更多的可调变量(各权值、阀值), 故网络可以以更复杂的方式逼近系统的外部特征. BP 模型的不足之处在于存储于各权上的知识人们无法理解, 所建立的模型难以用解析方式表达出来

改进方向:

① 样本集如何处理, 更能改善网络的学习效果, 提高识别精度;

② 研究网络的结构及诸参数与分类效果的关系;

③ 如何根据样本集的选择网络学习参数, 以提高网络的收敛速度;

④ 研究适用于分类识别问题的神经网络的闭环结构, 利用反馈信息, 提高网络预测的精度

参考文献:

[1] 王永骥, 徐 建. 神经网络控制. 机械工业出版社, 1998.

[2] Funahashi K J. the Appronimate Realization of Continuous Mapping by Neural Networks. Neutral Networks, 1989, (2).

[3] Rumelhart D E, Moclell J L. Parallel Distributed Processing: Exploration in the Microstructure of cognition. MIT Press, London, 1996.

[4] 袁曾任. 人工神经网络及其应用. 清华大学出版社, 1999.

[5] 陈 明. 神经网络模型. 大连理工大学出版社, 1995.

[6] 楼顺天, 施 阳. 基于MATLAB 的系统分析与设计——神经网络. 西安电子科技大学出版社, 1999.

[7] 王士同, 陈剑天. 问题求解的人工智能神经网络方法. 气象出版社, 1995.

[8] 胡守仁. 神经网络应用技术. 国防科技大学出版社, 1993.

A Model for DNA Sequence Clustering Problem

FENG Tao, KANG Zhe-wen, HAN Xiao-jun

(Dalian University of Technology, Dalian 116024)

Abstract: This paper presents a method applying artificial neural network (NN) to DNA clustering problem. First we use the probability statistics method to extract the characters from the 20 artificial DNA sequences whose categories are known. Thus we can get the character vectors of the DNA sequences and input them as samples into BP neuron NN for learning. We

employ the BP (back propagation) algorithm to train NN by use of the Neural Network Toolbox in MATLAB software package. In this paper, two three-story NN are created to input the extracted DNA character vectors as samples into them. After the training, characters are extracted from the 20 unclassified artificial sequence samples and 182 natural sequence samples to form the character vectors as input of the two NN for clustering. The results shows: the clustering method presented in this paper can classify the DNA sequences in quite high accuracy and precision. It is quite feasible to apply the artificial neural network to DNA sequence clustering.

DNA 分 类 模 型

杨 健, 王 驰, 杨 勇
指导老师: 王 鸣

(北京大学, 北京 100871)

编者按: 本文将DNA 序列的碱基的组合看作"文章"的关键词, 用逐步优选法对关键词进行优选并用分层分类的方法进行分类. 从理论上说, 这一方法可以提取较好的特征, 而且分类也较精细. 这一模型有一定创造性, 分析问题比较精细而贴近实际, 思路清楚, 叙述通顺简练.

摘要: 本模型充分利用了所给数据的特点, 运用统计、最优化等数学方法, 从已知样本序列中提炼出能较好代表两类特征的关键字符串, 据此提出量化的分类标准, 能较好的对任给DNA 序列进行分类. 首先, 从已知样本序列中用广度优先法选出所有重复出现的字符串, 并计算其标准化频率及分散度. 然后, 利用样本数据结合最小二乘法确定两类字符串各自的优先级函数, 并且逐步优化其参数使之达到稳定, 提高了可信度. 最后, 根据优先级函数找出关键词, 然后确定权数, 用层次分析法对未知样本进行分类, 并定出显著水平, 从而得到了一个比较通用的分类方法. 经过检验, 此方法对 21—40 号待测样本进行了很好的分类, 对后面的 182 个DNA 序列进行同样的操作, 也有较好的效果.

1 问题的重述(略)

2 模型假设

(1) 假定待分类样本 21—40 中既不属于A 类也不属于B 类的样本百分比不超过 5%.
(2) 假设 keyword 的重要性与 t 和 s 有确定的关系, 且只与 t 和 s 有关(t, s 定义见下).
(3) 假设不代表A、B 类特征的字符串在DNA 序列中是均匀分布的.

3 模型的分析

从所给的DNA 序列观察发现, 很多字符串重复出现的频率很高, 而且有些字符串在A 类和B 类中出现的次数有很明显的差距, 这暗示把某些字符串作为A, B 两类的一个分类标准. 所以应对A、B 两类已知样本做统计分析, 找出其中可能代表该类特征的字符串. 因为每个字符串重要性可能不一样, 所以对这些字串的重要性排序, 选出最能代表该类特征的一部分字串. 然后用这些字串作为标准判断验证A, B 两类, 看所选的标准的准确性, 最后用于任何一个DNA 序列的分类.

第 31 卷第 1 期 数学的实践与认识 Vol. 31 No. 1
2001 年 1 月 MATHEMATICS IN PRACTICE AND THEORY Jan. 2001

DNA 序列的分类模型

汤诗杰, 周 亮, 王晓玲
指导老师: 孙广中

(中国科技大学, 合肥 230026)

编者按: 本文提出了DNA 序列分类的三种模型, 其一, 基于A、G、T、C 四种碱基出现的频率; 其二利用了同一碱基在序列中的间隔, 这一信息是单纯考虑频率所不能包含的; 在第三种模型中, 作者把DNA 序列视为一个信息流, 考虑每增加一个字符所带来的信息增量. 尽管文中信息量的定义方式仍可讨论, 但本文思想新颖活跃, 有其独特之处. 本文最后的分类方法, 是以上三种的综合使用.

摘要: 本文针对DNA 序列分类这个实际问题, 提出了相应的数学模型. 为了很好的体现DNA 序列的局部性和全局性的特征, 我们给出了衡量分类方法优劣的标准, 即在满足一定限制条件的情况下, 是否能充分反映序列的各方面特性.

依据我们提出的判别标准, 单一标准的分类是无法满足要求的. 我们的方法是侧重点不同的三种方法的综合集成. 这三种方法分别体现了序列中元素出现的概率, 序列中元素出现的周期性, 序列所带有的信息含量. 利用这个方法, 完成了对未知类型的人工序列及自然序列的分类工作. 最后, 对分类模型的优缺点进行了分析, 并就模型的推广作了讨论.

1 问题的提出(略)

2 问题的分析

这是一个比较典型的分类问题, 为了表述的严格和方便, 我们用数学的方法来重述这个问题. 已知字母序列 $S_1, S_2, S_3 \cdots\cdots S_{40}$, $S_i = x_{i1}x_{i2}x_{i3}\cdots x_{in_i}$, 其中 $x_j \in \{a, t, c, g\}$; 有字符序列集合 A, B, 满足 $A \cap B = \phi$, 并当 $1 \le i \le 10$ 时, $S_i \in A$; 当 $11 \le i \le 20$ 时, $S_i \in B$. 现要求考虑当 $21 \le i \le 40$ 时, S_i 与集合 A 及集合 B 的关系.

在这里, 问题的关键就是要从已知的分好类的 20 个字母序列中提取用于分类的特征. 知道了这些特征, 我们就可以比较容易的对那些未标明类型的序列进行分类. 下面我们将首先对用于分类的标准问题进行必要的讨论.

3 分类的标准及评价

首先, 我们提取的特征应该满足以下两个条件:

(1) 所取特征必须可以标志A 组和B 组. 也就是说, 我们利用这些特征应该可以很好的区分已经标示分类的 20 个序列. 这是比较显然的一个理由.

(2) 所取特征必须是有一定的实际意义的. 这一点是决不能被忽视的. 比如, 如果不考虑模型的实际意义, 我们就可以以序列的开头字母为分类标准: 已知在B 类中的十个序列都是以 gt 开始的, 而已知在A 类中 10 个序列没有以 gt 开始的, 甚至以 g 开始的都没有. 显然这是满足上面的第一个条件的. 如果仅因此就认为这种特征是主要的, 并简单的利用这个特征将所有待分类的序列分成两类, 显然是不甚合理的.